TÉCNICAS DE DATA MINING APLICADAS A DATOS DE TRANSPORTE PÚBLICO

Juan Carlos Otaegui. Departamento de Ingeniería. UNLAM
Cristóbal R. Santa María. Departamento de Ingeniería. UNLAM
Florencio Varela 1903 San Justo Pcia. de Buenos Aires
54-011-44808952
juancarlosotaegui@yahoo.com.ar
csantamaria@unlam.edu.ar

RESUMEN

El objetivo general es aplicar técnicas de data mining en el análisis de la información digital que ya hoy producen, o bien que podrían producir los medios de transporte público con el fin de obtener nuevas perspectivas que colaboren exponiendo patrones de uso desconocidos. Generar nueva información en el campo por medio del análisis inteligente de bases de datos de gran tamaño proporciona una herramienta importante para la toma de decisiones que mejoran la experiencia del usuario de transporte público en las grandes ciudades.

Se han realizado con éxito investigaciones en distintas ciudades del mundo que disponen de información digitalizada en grandes bases de datos. El denominado Sistema Único de Boleto Electrónico (SUBE) es un medio para abonar con una sola tarjeta viajes en colectivos, subtes y trenes adheridos, en la Región Metropolitana de Buenos Aires. Fue implementado durante 2011, cuenta ya con más de un millón de tarjetas emitidas y continúa extendiéndose a las principales ciudades del país. Esto permitirá contar con una base de datos lista para ser explotada con técnicas de data mining tales como clustering de usuarios, arboles de decisión y modelos que predigan con precisión las rutas optimas y tiempos de viaje entre otras aplicaciones posibles.

La base de datos de viajes que genera SUBE junto con las características de los usuarios puede ser totalmente complementada con la tecnología GPS (Global Positioning System), disponible en gran parte de los vehículos, y también con el estado siempre dinámico de las vías, avenidas, calles, autopistas, semáforos, pronósticos del tiempo y una larga lista de bases de datos que pueden colaborar para realizar modelos predictivos de comportamiento y permitir así la toma de decisiones óptimas.

CONTEXTO

La línea de trabajo que aquí se presenta se inscribe en el proyecto de investigación de técnicas de minería de datos aplicadas sobre bases de datos generadas a través de las transacciones que generan las denominadas Smart Cards con la cual los usuarios registran sus viajes en el uso del transporte público. El proyecto tiene la finalidad de estudiar la aplicación de técnicas de data mining como clustering de usuarios, arboles de decisión y generación de modelos para hallar patrones que clasifiquen y predigan comportamientos de usuarios

INTRODUCCIÓN

La mejora constante de la calidad del transporte público debe ser un objetivo primordial en las grandes ciudades del mundo donde se concentran la mayor cantidad de viajes y donde cada vez se hace más evidente una necesidad de utilizar las últimas tecnologías disponibles para lograr este objetivo que tanto influye en la calidad de vida de la población.

Las denominadas Smart Cards o Tarjetas magnéticas asociadas a un usuario con la cual registra y abona sus viajes se utilizan en las principales ciudades del mundo como Seul, Chicago, Barcelona, Madrid, Londres, etc. Generalmente, estas Smart Cards, forman parte de una red integrada para utilizar en varios medios de transporte como buses, trenes, o subterráneos en una amplia zona metropolitana. Estas tarjetas permiten registrar digitalmente los viajes en bases de datos de forma que se pueda generar un repositorio con datos básicos de cada viaje como n-upla "fecha/hora, usuario, parada, monto abonado".

Por otra parte, otra tecnología utilizada en la actualidad en los transportes públicos es la información proporcionada por los GPS que están hoy instalados en una gran parte de las unidades. Este sistema permite conocer el posicionamiento exacto en un momento determinado de una unidad en su recorrido y almacenarlo en un log o bien transmitirlo inmediatamente a una central con una frecuencia de tiempo determinada. Toda esta información geo referenciada asociada a trayectos fijos definidos en el transporte público representan otra base de datos de la cual se puede extraer el estado diario con el detalle a nivel de hora y minuto de las arterias por las que circulan las unidades en el caso de los buses o retrasos en la frecuencia para los trenes o subterráneos.

Durante la investigación de transporte urbano público de Buenos Aires (INTRUPUBA) realizada en 2007 se pudo recabar información geo referenciada con tecnología GPS en toda el área metropolitana agregando además información de cuenta de personas en colectivos que agrega ascenso y descenso. Esto último es muy importante ya que se pueden conmensurar los destinos reales de los pasajeros.

Durante 2011 se implementó el Sistema Único de Boleto Electrónico (SUBE) en Buenos Aires y cada vez son más las ciudades que se incorporaran al sistema nacional que permite digitalizar la información de los viajes.

El posible repositorio a analizar es una fuente de conocimiento que se puede abordar con distintas técnicas de análisis inteligente de la información y algoritmos empleados en grandes bases de datos y como resultado se puede esperar conocer más sobre el comportamiento general y particular del uso del transporte público. Este trabajo pretende hacer un recorrido del estado del arte en lo que respecta a investigaciones que se han realizado en los repositorios disponibles con tarjetas inteligentes o boletos electrónicos y aplicación de tecnología GPS como soporte para la toma de decisiones y mejora de la calidad del transporte público. A la vez intenta establecer algunas ideas respecto del proceso de los datos generados localmente.

LÍNEAS DE INVESTIGACIÓN Y DESARROLLO

Se propone que tal estado del arte contribuya a diseñar un Data Warehouse que permita aplicar técnicas de clustering para obtener grupos de usuarios según diversas dimensiones disponibles recolectadas por SUBE como, por ejemplo, horarios, edades, orígenes y destinos de los viajes, categoría de la tarifa, datos de las unidades de transporte. Se plantea también sugerir nuevas dimensiones para caracterizar cada viaje como la frecuencia por usuario del mismo, los componentes externos de contexto como estado del tránsito y las alternativas disponibles. Con cada cluster de usuarios se pretende predecir el tiempo de viaje estimado y medir la diferencia respecto a los tiempos de viajes publicados estimados. De esta forma se

podrá conocer el grado de desviación para cada uno de los grupos resultantes..

RESULTADOS Y OBJETIVOS

El trabajo ha comenzado relevando las distintas investigaciones que se realizaron sobre bases de datos que contienen información análoga a la disponible en SUBE**.** Se analizaron investigaciones similares sobre la utilización de los subtes en Londres, los buses o colectivos en Quebec y los tranvías en Lisboa. También se consideraron trabajos realizados en Turín y Singapur.

En la ciudad de Londres se analizó un conjunto de 89 millones de viajes (70 % de viajes en subte y el resto en tren) realizados durante los 31 días de marzo de 2010. Así se identifican varios grupos de usuarios y se describió la característica de cada uno de ellos, horarios y días de la semana en que se utilizaba la red. Se observó que los tiempos invertidos en viaje no son uniformes para todos los usuarios y tienen alta correlación con la familiarización del trayecto. Además se constató que los tiempos de viaje publicados difieren mucho de los tiempos de viajes reales. Se estableció también que es posible producir modelos personalizados para los usuarios que permitan conocer de una manera precisa cuanto tiempo demorara un viaje **(1)**.

En Quebec se estudiaron 6.2 millones de registros de viajes en buses para un periodo comprendido entre 01/01/2005 y 01/10/2005 realizados por portadores de la tarjeta electrónica. Se identificaron varios grupos de usuarios y se describió la característica de cada uno de ellos en cuanto a frecuencia, horarios y días de la semana en que se utiliza la red **(2)**.

En Lisboa se analizaron logs de buses y tranvías de los meses diciembre 2009, enero, febrero y marzo de 2010. Cada log mensual contuvo 390.000 observaciones. Se registró con tecnología GPS el detalle del momento exacto en que la unidad paso por cada una de las paradas. La idea de tal trabajo fue realizar mediante la información histórica y la generada en tiempo real, una predicción del tiempo estimado de un recorrido para que los usuarios dispusieran de la información a través de una plataforma online y pudieran decidir que trayectos, combinaciones y líneas escoger. Para ello se aplicó una combinación de árboles de decisión y regresión lineal que generaron un modelo predictivo del tiempo de duración del viaje **(3)**.

Una investigación de parecido tipo se llevó a cabo en Turín sobre 123.000 observaciones de una encuesta pública referida a atributos socioeconómicos y modos de transporte que se utilizan Se aplicaron Reglas de Asociación para lo cual las variables fueron re codificadas a dummies. Se estudiaron reglas con soporte mayor a 0,01 y un lift de 1,1. A través de tales reglas se pudo interpretar que existe una sustitución del vehículo propio por medio de transporte público mientras que no existe una sustitución del vehículo/transporte público por un medio no motorizado. Se estima que esto tiene que ver con la falta de infraestructura que facilite dicho reemplazo **(4)**. Se destaca que un proceder similar podría seguirse a partir de los datos de INTRUPUBA.

En Singapur a través de los datos de viajes registrados por tarjetas electrónicas se realizaron caracterizaciones de las actividades de los usuarios. Se analizaron los horarios en que se realizaban los viajes, los destinos según el tipo de actividad laboral que se efectuaba en cada uno de ellos, para realizar un cruce con los datos relevados en una encuesta nacional de transporte del año 2008. Se incorporó así, como una nueva dimensión, la posibilidad de identificar cambios en la actividad laboral de los

usuarios basándose en los registros de viaje **(5)**.
Una vez establecido el estado del arte se comenzó a analizar como podría complementarse un Datawarehouse con base de datos externas con la información proporcionada por GPS, los servicios de información meteorológica o de estado del tránsito en general.
La viabilidad del proyecto completo del trabajo sobre los datos del AMBA está dada por contar con una vasta red que contiene a 7 líneas de ferrocarriles, 6 subterráneos y más de 340 líneas de colectivos, siendo estos últimos el medio más utilizado con el 75% del total de los viajes realizados. La tarjeta SUBE permite digitalizar e individualizar los viajes registrados así como también la características de los mismos con una base de más de 2300 millones viajes al año **(6)**.
De acuerdo a los estudios detallados se propone utilizar la información de SUBE para realizar un análisis global de comportamiento de los usuarios. Luego de esto realizar un agrupamiento en clusters para después exponer las diferencias entre el comportamiento global y el que tienen cada uno de los clusters de usuarios. Evaluar en este caso la actividad por hora, el tiempo promedio de viaje, el promedio de cantidad de viajes por mes-usuario, la cantidad de usuarios que repiten viaje, loops (inicio-destinos-vuelve al inicio), cantidad de estaciones visitadas. etc. El clustering se aplicará sobre un conjunto de datos que incluye usuario, hora categorizada en 5 segmentos a determinar (por ejemplo: primera mañana, mañana hora pico, mediodía, tarde, noche) y cantidades de ocurrencias en el mes de estudio. La medida de distancia entre los usuarios se evaluará como el módulo de la resta entre cada vector. A menor valor resultante mayor parecido entre los usuarios. Se seleccionará el mínimo valor en la matriz y se crea un nuevo cluster donde el valor del centroide se calcule como la media de los valores de los integrantes del cluster. Este proceso se repite hasta que no exista una distancia menor a un umbral a determinar según los clusters obtenidos y la interpretación funcional de los mismos.
Luego se generarán modelos predictivos en tiempos de viajes para cada uno de los viajes resultantes teniendo en cuenta las los siguientes puntos:

- Siempre se esperará el camino más corto calculado con el algoritmo de Dijkstra.
- Tiempo medio real (no el publicado) entre estaciones y combinaciones.
- Contexto horario
- Familiaridad con el recorrido.

Se pretende así conocer y descubrir patrones de viajes nuevos en cada uno de los clusters que se obtengan y modelos que predigan tiempo de viaje para cada uno. Esta información puede proveer múltiples usos fundamentales como herramienta para la toma de decisiones en la materia.
En cuanto al diseño del data warehouse, mediante la información disponible se propone un posible modelo dimensional que combina algunas dimensiones y hechos relacionados con el sistema de información transaccional y fuentes de datos con información geo-referenciada.
Algunos ejemplos de nuevas dimensiones que se podrían generar son:

- Cluster de usuarios según el uso del transporte: Usuarios lunes a viernes, fines de semana, usuario ocasional, según la franja horaria en la que utiliza el servicio, etc.
- Cluster de trayectos según su asociación con otros trayectos: Relación entre transportes,

trayectos opcionales, distancia entre paradas, etc.
- Estado del tránsito: registrar el estado de las arterias de acuerdo a como se registran los viajes, estado meteorológico, arteria en reparación, huelgas, etc.
- Incorporar base de datos externas: parque automotor, información choferes, información de semáforos, pronósticos meteorológicos, pronósticos informativos relacionados con tránsito, frecuencias, etc.

Estas múltiples dimensiones y hechos pueden ofrecer una herramienta de análisis para:

- Realizar reportes operacionales con datos históricos.
- Realizar tableros de control.
- Realizar análisis ad hoc (no predefinidos) a través de tecnología OLAP que permitan a los analistas construir sus propios análisis dinámicamente.
- Realizar predicciones creando modelos.
- Disponer de un "Active Datawarehouse": Este tipo de DW permite tener la información casi on-line para tomar decisiones en el momento. Ejemplo detectar casos de contingencia rápidamente para notificar y reasignar rutas opcionales.

BIBLIOGRAFÍA

1 - Neal Lathia , Chris Smith, Jon Froehlich , Licia Capra. Individuals among commuters Building personalised transport information services from fare collection systems www.elsevier.com/locate/pmc 2012

2 - Martin Trepaniera, Bruno Agard. Measuring Transit use variability with smat-card data. Transport Policy 14 193–203. 2007

3 - David Alves, Luis M. Martinez, José M. Viegas Retrieving real-time information to users in public transport networks: an application to the Lisbon Bus System Procedia - Social and Behavioral Sciences 54 .470 – 482. 2012

4 - Marco Diana. Studying Patterns of use of transport modes through data mining application to U.S. National Survey dataset. http://dx.doi.org/10.3141/2308-01 2012

5- Sun, L, Lee, DH, Erath, A, Huang, X. Using Smart Card Data to Extract Passenger's Spatio-temporal Density and Train's Trajectory of MRTSystem

6- INTRUPUBA - Investigación del Transporte Urbano Público de Buenos Aires. Ministerio Planificación Federal. Secretaria de Transporte. 2007

7- Paul Bouman, Milan Lovric, Ting Li, Evelien van der Hurk, Leo Kroon, Peter Vervest. Recognizing Demand Patterns from Smart Card Data for Agent-Based Micro-simulation of Public Transport.